Prompting for the Unknown: Leveraging In-Context-Learning for Few-Shot Open Set Classification

Alexander Grote
Karlsruhe Institute of Technology
alexander.grote@kit.edu

Anuja Hariharan
Karlsruhe Institute of Technology
anuja.hariharan@kit.edu

Christof Weinhardt
Karlsruhe Institute of Technology
weinhardt@kit.edu

Abstract

Recognising customer intent is crucial for applications such as chatbots and virtual assistants, requiring accurate interpretation of user inputs. While traditional intent recognition systems depend on large datasets and complex machine learning pipelines, large language models (LLMs) offer competitive performance with significantly less training data through in-context learning (ICL). In this work, we assess the effectiveness of ICL for intent recognition, with a particular focus on detecting out-of-distribution (OOD) inputs. We explore prompting strategies to improve OOD detection and systematically evaluate few-shot classifiers under varying OOD proportions. Our results show that implicit prompting strategies yield better precision for OOD detection, while explicit strategies excel at recall. Moreover, we confirm that LLMs perform comparably to conventional classifiers on in-distribution data. However, a significant fraction of OOD errors are non-overlapping between LLMs and traditional models, highlighting limitations in LLM robustness and suggesting new directions for enhancing generalisation in intent recognition systems.

Keywords: In Context Learning, Text Classification, Out-Of-Distribution Detection

1. Introduction

Customer intent recognition plays a critical role in the digital transformation of businesses, particularly in applications like chatbots and virtual assistants, where an accurate understanding of user goals directly impacts service quality and customer satisfaction (Kuligowska and Kowalczuk, 2024; Sidlauskiene et al., 2023; Suryanto et al., 2023). Traditionally, these systems have relied on discriminative machine learning approaches, such as IntentBert (Zhang et al., 2021) or dual sentence encoders (Casanueva et al., 2020). All of these approaches require substantial annotated data and engineering effort (Ran et al., 2024). However, the advent of instruction-fine-tuned Large Language Models (LLMs), such as ChatGPT (OpenAI, 2022), Claude Sonnet (Anthropic, 2023), and Gemini (Gemini Team, 2023), has led to a surge in the adoption of LLM-based solutions for intent recognition tasks (Arora et al., 2024).

While LLMs have demonstrated strong performance across a wide range of domains, real-world intent classification remains challenging. In practice, systems must often deal with ambiguous queries (He and Garner, 2023), evolving user intents (Ran et al., 2024), unknown categories (P. Wang et al., 2024), and scenarios where only minimal labelled data is available (Arora et al., 2024). These constraints have sparked interest in few-shot open set recognition, a setup in which models are expected to generalise to both seen and unseen intent classes using only a handful of examples (L.-Y. Sun and Chu, 2024). Fine-tuning LLMs for such use cases is often impractical due to the need for domain-specific expertise (Jin et al., 2024), high-quality annotated data (J. Sun et al., 2024), and computationally expensive infrastructure (Oliver and Wang, 2024). As a result, practitioners increasingly rely on In-Context Learning (ICL), where models learn from a few labelled examples presented in the prompt (Dong et al., 2024). For instance, in a travel agency setting, the prompt might show just one example of ”paying fee” and ”booking flight”, and the LLM must then classify a new query, such as ”I need to change my flight” correctly. However, if a user instead asks about a bank statement, this falls outside the known intent categories and becomes an

URI: https://hdl.handle.net/10125/111547
978-0-9981331-9-5

out-of-distribution (OOD) case.

Despite the low entry barrier and flexibility of ICL-based approaches, several open challenges remain. One major concern is how LLMs handle OOD inputs, and how we can improve the knowledge transfer from in-distribution data to OOD data (P. Wang et al., 2024). While foundational work has shown LLMs exhibit strong zero-shot (Zhu et al., 2024) and few-shot (Brown et al., 2020) capabilities, it remains unclear how to prompt them for robust OOD detection. One strategy is to include explicit OOD examples in the prompt. However, this approach might force the LLM to override its semantic priors and conceptualise the unknowns as another class defined by the given examples (Kossen et al., 2023). Another open question is whether OOD detection should be treated as a separate binary classification task, which requires two prompts for the final classification. Yet, most recent works (Arora et al., 2024; P. Wang et al., 2024) have only considered a single prompt solution, incorporating the classification and OOD detection simultaneously. The trade-offs between these approaches remain underexplored.

From a practitioner's perspective, a key decision is whether to rely on traditional intent classification methods or to adopt LLM-based prompting techniques. In this paper, we explore these trade-offs in the context of real-world applications characterised by limited labelled data (Baier et al., 2019) and the presence of unseen intents (G. Chen et al., 2024). Despite related work showing that ICL can be approximated with linear, gradient-descent-based models (S. Li et al., 2024; Von Oswald et al., 2023), it remains unclear how correlated the predictions between traditional machine learning (TML) and LLMs are in real-world applications.

Given these motivations, we pose the following research questions:

- **RQ1**: Prompting Strategies

  Which prompting methods strategies (explicit vs. implicit, one-shot vs. few-shot) are most effective for identifying unknown classes in prompt-based OOD with LLMs?

- **RQ2**: TML versus LLMs

  a) How does the classification robustness of few-shot LLM-based classifiers compare to traditional models as dataset openness increases?

  b) Do their classification error patterns differ?

To investigate our research questions, we conduct an empirical study[1] using three established intent classification datasets and simulate four open set scenarios, varying the number of unknown classes from 0 to 60 % during inference time. In terms of models, we use four state-of-the-art LLMs and three established baseline few-shot classifiers. This setup allows us to robustly test different prompting strategies that aim to reject unknown data points as OOD while also classifying known classes. It also enables us to compare their robustness across different difficulty levels with existing machine learning classifiers and analyse if their errors are correlated.

Our results on various prompting strategies (RQ1) reveal that the choice between implicit and explicit prompting should be guided by whether recall or precision is more critical in the given OOD detection task. Specifically, explicit prompting tends to favour higher recall, whereas implicit prompting yields better precision. Furthermore, when compared with traditional few-shot classifiers, the LLMs yield a significant number of non-overlapping errors on OOD detection tasks. This suggests that while LLMs and traditional classifiers may achieve comparable overall performance, they fail on different data points, highlighting the complementary nature of these models in OOD scenarios.

With our analysis, we reinforce the general conclusions of prior work by P. Wang et al. (2024) and Arora et al. (2024) that ICL can match or even exceed the performance of conventional machine learning models on intent recognition tasks. Our contribution extends this line of research by providing a focused examination of OOD detection, specifically through a comparative analysis of prompting strategies, offering actionable insights for prompt selection. Unlike recent work, we also analyse error patterns between traditional few-shot classifiers and LLMs, highlighting the strengths and weaknesses of ICL. These findings are not only relevant to our current benchmark but are also generalizable to similar classification and detection tasks, such as intent detection under domain shift (Lang et al., 2024), and content moderation (Kumar et al., 2024), where rare or unseen categories can bypass a content moderation filter.

The remainder of this paper is structured as follows: Section 2 introduces related work on ICL and open intent recognition. In the subsequent Section 3, we outline the methodology of our empirical study, including the datasets, models, prompting strategies, and the study design. We then describe the results in Section 4 and discuss them in Section 5. Finally, we conclude the paper and provide a brief outlook of our research in Section 6.

[1] The entire codebase to reproduce the experiments is available at https://github.com/alexandergrote/llm_osr.

2. Related Work

2.1. In-Context-Learning

Brown et al. (2020) defines ICL as a new learning paradigm for LLMs to learn from given examples. While LLMs inherently exhibit robust ICL capabilities (Dong et al., 2024), their effectiveness can be further enhanced. Dong et al. (2024) broadly distinguish two primary possibilities for improving ICL: (1) model training and (2) prompting techniques. In terms of model training, incorporating task-specific context into pretraining data has proven beneficial (Gu et al., 2023; Shi et al., 2023). Since pretraining data often lacks alignment with instruction formats, introducing a warm-up phase between pretraining and inference improves ICL performance (M. Chen et al., 2022). To enhance input-output mappings, Wei et al. (2023) propose symbolic learning, where labels are replaced with symbolic representations. Additional fine-tuning on instruction-based data also yields performance gains (Min et al., 2022; X. Wang et al., 2023). When it comes to prompting techniques, using semantically similar examples improves performance (J. Liu et al., 2022), and the order of examples also plays a role (Lu et al., 2022). Detailed task instructions further boost results, especially for complex reasoning tasks (Dong et al., 2024), and LLMs can generate such instructions autonomously (Honovich et al., 2023). Similar to elaborated instructions, Chain-of-Thought (CoT) prompting, which guides models through intermediate reasoning steps, is also known to increase performance (Wei et al., 2022).

In our work, we follow best practices of ICL, such as CoT and apply them to classification problems on three datasets. By comparing the classification performance with traditional few-shot classifiers, we seek to enrich the understanding of how ICL works and where it has its strengths and weaknesses.

2.2. Open Intent Recognition

In this work, we investigate the application of LLMs to open intent classification, a task that involves both (1) detecting unknown intents and (2) classifying known ones, and has also been widely studied. Traditional few-shot learning approaches, such as Prototypical Networks (Snell et al., 2017), Matching Networks (Vinyals et al., 2016), and more recent metric-learning based classifiers like ContrastNet (J. Chen et al., 2024), have been effectively applied in settings with scarce data. Arora et al. (2024) compare SetFit (Tunstall et al., 2022), a representative few-shot machine learning classifier, with several LLMs, and additionally propose a hybrid framework that balances latency and accuracy. Parikh et al. (2023) study how we can leverage domain information and data augmentation for enhanced intent recognition with GPT-3. P. Wang et al. (2024) also evaluate LLMs, specifically ChatGPT, against contrastive learning-based classifiers, considering various levels of openness in the task. Meanwhile, B. Liu et al. (2024) focus on fine-tuning LLMs to improve OOD detection, a strategy which Arora et al. (2024) argue is impractical for real-world deployment due to its resource demands.

Compared to existing work, we not only confirm the effectiveness of LLMs in intent classification but also contribute novel insights by comparing two alternative prompting strategies. Furthermore, we analyse the error patterns between traditional few-shot classifiers and ICL with LLMs, revealing promising directions for future research.

3. Methodology

3.1. Datasets

We use three well-established datasets for our empirical study: 1) BANKING77 (Casanueva et al., 2020), 2) CLINC150 (Larson et al., 2019) and 3) HWU64 (X. Liu et al., 2021). All three datasets represent intent classification problems with a high number of classes and rather short texts. Table 1 summarises statistics about the datasets and provides one exemplary data point each. CLINC150[2] is the largest dataset in terms of both total samples (n=22,500) and number of classes (n=150), with a uniform class distribution. In contrast, HWU64 and BANKING77 have fewer classes and more variation in the number of samples per class, with BANKING77 showing the longest average text length (11.7 words) among the three.

3.2. Models

To derive meaningful insights into ICL, we select four well-established LLMs, representing a range of model sizes. These include **Llama 3.3** (70B parameters; Meta AI, 2024b), **Gemma3** (27B; Kamath et al., 2025), **Phi-4** (14B; Abdin et al., 2024), and **Llama 3.1** (8B; Meta AI, 2024a). To obtain the maximally deterministic output for a classification problem, we have used a temperature value of 0.

[2] To ensure compatibility with our data sampling strategy, we excluded the original OOD samples provided in the dataset. These samples are drawn from a distinct distribution not aligned with our controlled setup. Instead, we generate our own OOD samples based on intent labels, ensuring consistency across all datasets and greater control over the OOD evaluation.

Table 1: Dataset statistics and examples.

Dataset	Samples		Per-Class Samples			Text Length (Words)		Example
	Total	Classes	Min	Max	Avg	Avg	Range	
CLINC150	22,500	150	150	150	150.0	8.3	2–136	The sound is too low → change_volume
HWU64	11,106	64	39	197	173.5	6.6	2–133	I need you to set an alarm → alarm_set
BANKING77	13,083	77	75	227	169.9	11.7	13–433	My card is due to expire soon. → card_about_to_expire

To compare the classification performance of ICL with established few-shot machine learning classifiers, we use **SimpleShot** (Y. Wang et al., 2019), **FastFit** (Yehudai and Bandel, 2024), a computationally more efficient version of SetFit (Tunstall et al., 2022), and **ContrastNet** (J. Chen et al., 2022). While the former represents a nearest class mean classifier, the remaining algorithms utilise contrastive learning to maximise their classification accuracy. To adapt the classifiers for an open set scenario, we use their probability estimates and learn a threshold that rejects data points as unknown if the algorithm is uncertain about their classification. We determine this threshold as part of our hyperparameter tuning with the Tree-Structured-Parzen Estimator (Bergstra et al., 2011) on the validation dataset.

3.3. OOD Prompting

In our work, we distinguish between two prompting strategies for detecting OOD samples. The first is explicit prompting, where the prompt directly addresses the presence of OOD instances. The second is implicit prompting, where the model is guided to identify OOD samples without directly referencing them. Figure 1 provides examples of both strategies. Additionally, to help the LLM understand unknown classes, we either provide examples of such classes from our validation dataset, which we refer to as ”few-shot”, or we provide no such examples, which we refer to as ”zero-shot”.

Regardless of the prompting strategy, we select only the top 5 semantically similar data points as examples, which is known to enhance classification performance compared to random sampling (J. Liu et al., 2022). To create such a semantically similar set of data points, we use an angle-optimised embedding (X. Li and Li, 2024). In case of supplying examples of unknown classes, we use five random data points from the validation dataset whose classes are not contained in the train dataset. To avoid any ranking bias in our resulting selection, we shuffle the data points prior to passing them to the LLM (Mina et al., 2025). In addition, we also append the famous zero-shot CoT phrase ”Let’s think step by step” to the prompts above to encourage the LLM to think through its responses step-by-step (Kojima et al., 2022). The last component of the supplied prompt are JSON instructions, which enables us to process the data points in a structured manner. Within these JSON instructions, we ask the model to first explain its reasoning and then provide the class label.

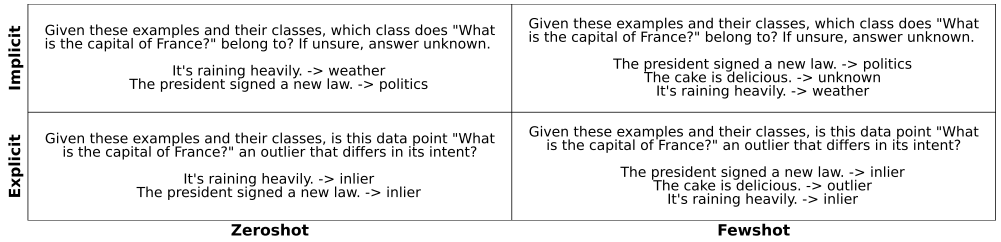
3.4. Open Set Data Split

In OSR problems, we assume that not all classes are available during the inference process, which affects the train, validation and test split. We follow the example of Geng et al. (2021) to create these data splits. First, we define the number of known and unknown classes by random selection. The number of the known classes θ thereby depends on a user-defined percentage. Based on this selection, we then proceed to define the test data set, which consists of 40 % of the data points assigned to known classes and all of the data points that belong to the unknown class. The remaining 60 % of the known data points comprise the training dataset, which we further divide into a fitting and validation set. Similar to the train and test split above, we also simulate an open set for the validation dataset. This time, we select $\frac{2}{3}\theta + 0.5$ as the number of known classes, which we again select randomly, and apply the same 60:40 split as before. To adapt the OSR datasplit to a few-shot setting, we randomly draw five examples per class from the fitting and validation set. Additionally, to manage cost and computational efficiency associated with the LLM REST API queries, we have limited the resulting test set to 5,000 randomly chosen data points. Despite this, our classification results remain consistent with those reported in prior work (Yehudai and Bandel, 2024), where unknown classes were not selected.

4. Results

4.1. OOD Prompting Strategies

To evaluate the impact of different prompting strategies on unknown detection performance, we report precision, recall, and the F1 score in Figure 2. We assign 20% of all classes as unknown and aggregate the classification outcomes across three datasets and four models by prompting strategy. Each



	Zeroshot	Fewshot
Implicit	Given these examples and their classes, which class does "What is the capital of France?" belong to? If unsure, answer unknown. It's raining heavily. -> weather The president signed a new law. -> politics	Given these examples and their classes, which class does "What is the capital of France?" belong to? If unsure, answer unknown. The president signed a new law. -> politics The cake is delicious. -> unknown It's raining heavily. -> weather
Explicit	Given these examples and their classes, is this data point "What is the capital of France?" an outlier that differs in its intent? It's raining heavily. -> inlier The president signed a new law. -> inlier	Given these examples and their classes, is this data point "What is the capital of France?" an outlier that differs in its intent? The president signed a new law. -> inlier The cake is delicious. -> outlier It's raining heavily. -> inlier

Figure 1: Conceptual overview of employed prompting strategies with arbitrary examples.

experimental configuration is repeated five times, resulting in a total of 60 data points per strategy. Since the data points associated with each strategy are not independent and the metric distributions deviate from normality, we use the Wilcoxon Signed-Rank Test (Wilcoxon, 1945) for non-parametric, pairwise statistical comparisons. Additionally, we report the Rank-Biserial Correlation[3] $r_{bc} \in [-1, 1]$ to quantify the effect size and consistency of the direction of observed differences (Kerby, 2014). The results indicate that, in terms of precision, the implicit prompting strategy outperforms the explicit prompting strategy with statistical significance ($p < 0.001$) and a large effect size $r_{bc} \approx 1$. However, in terms of recall and f1 score, the explicit zero-shot prompting strategy outperforms all the other strategies with statistical significance and strong consistency. Supplying examples of unknown classes has a medium-sized effect, though the absolute mean performance difference is small for implicit strategies. For explicit few-shot strategies, omitting unknown class examples leads to better precision, while including them improves recall and f1 scores.

4.2. Open Set Text Classification

To compare the overall classification results of LLMs with conventional few-shot classifiers, we look at the known and unknown classes separately. Figure 3 shows the mean f1 score across five different seeds for all models and varying proportions of unknown classes. The LLMs, prompted with an implicit zero-shot strategy, exhibit similar performance trends to existing few-shot classifiers, which is consistent across the known and unknown classes. However, the overall performance varies substantially from model to model. While the Llama 3.1 8B model exhibits the weakest performance among the LLMs, Phi-4 and Llama 3.3 70B are among the top-performing models. When it comes to the traditional models, no such clear tendencies between the models are visible. Instead, the performance depends significantly on the dataset and the number of unknown classes.

To further analyse the errors made by the models, we again differentiate between the known and unknown classes. However, to derive generalisable findings between traditional few-shot classifiers (TML), we take the majority vote of the LLM models versus the majority vote of the traditional models. The contingency tables in Figure 4 illustrate the distribution of correct and incorrect predictions across the two model classes. To quantify the degree of overlap in prediction errors, we compute the Phi coefficient and perform chi-square tests to assess statistical significance. For the known classes, we observe a strong error overlap with a Phi coefficient of $\Phi = 0.50$, indicating a substantial association between the models' prediction errors ($p < 0.001$). In contrast, for the unknown classes, the overlap is weaker, with a Phi coefficient of $\Phi = 0.14$, though still statistically significant ($p < 0.001$). These results demonstrate that while both models tend to make similar predictions on known classes, their errors on unknown classes diverge more substantially, suggesting different generalisation behaviours in open set conditions.

5. Discussion

In this study, we have investigated the impact of implicit and explicit prompting strategies on the OOD detection (RQ1) and the overall performance of LLMs compared to existing few-shot classifier models (RQ2). We have found that implicit prompting yields better precision, while explicit prompting yields higher recall. A possible explanation for the latter effect is that, under a Bayesian view of ICL (Dong et al., 2024), explicitly framing outlier detection in the prompt makes the model more likely to predict the unknown class. Additionally, by clearly focusing the wording on identifying what makes an example different, the LLM may explicitly search for such patterns and thus, be more likely to recognise it as unknown. These two effects produce more false positives and therefore

[3] According to Cohen (1988), an effect size $|r_{bc}| \geq 0.465$ is considered large, $|r_{bc}| \in [0.304, 0.465)$ medium, and $|r_{bc}| \in [0.125, 0.304)$ small. An effect size $|r_{bc}| < 0.125$ is negligible.

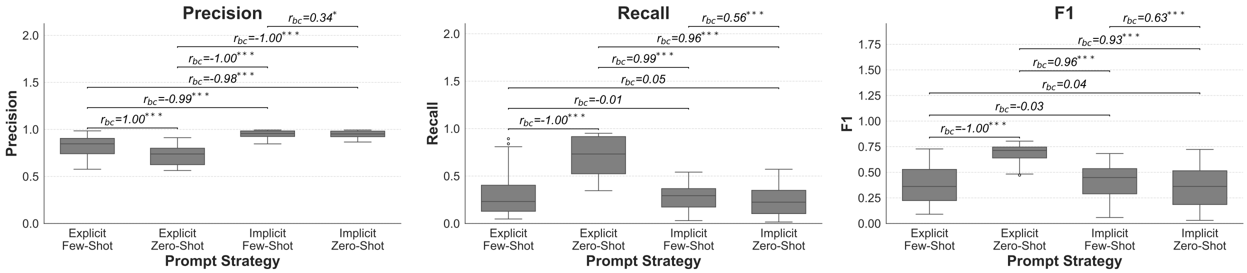
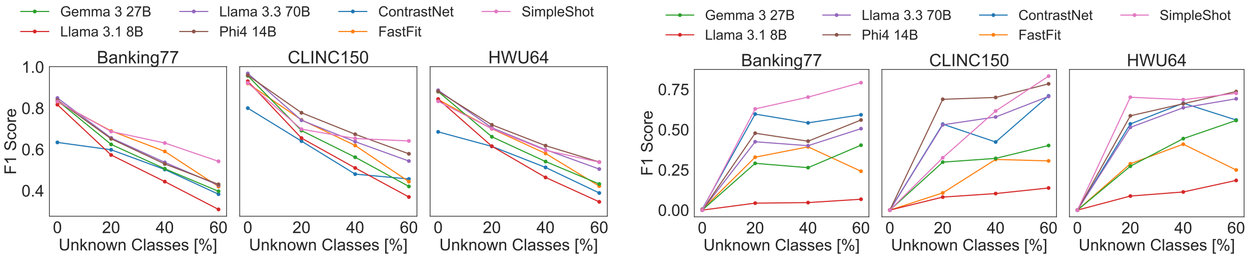


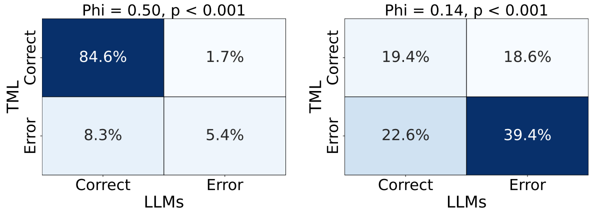
Figure 2: Classification performance for unknown detection under each prompting strategy. Statistical significance is indicated as follows: *** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$.



(a) F1 score on known classes.

(b) F1 score on unknown classes.

Figure 3: Comparison of classification results across varying degrees of unknown classes.



(a) Known classes.

(b) Unknown classes.

Figure 4: Contingency tables showing error overlap between few-shot classifiers (TML) and LLMs.

reduce precision, whilst increasing recall. Implicit prompting, on the other hand, treats the unknown class as one of many. As a result, the prompt does not prime as much as explicit strategies for unknown detection, leading to fewer unknown predictions. This reduction suggests that implicit strategies are more conservative, predicting the unknown class primarily when there is a substantial disparity between the supplied examples. Consequently, they achieve higher precision. Comparing the performance of LLMs to traditional few-shot classifiers, we observe that LLMs follow similar trends as traditional classifiers, but often achieve better or comparable performance. However, it is essential to note that the performance of LLMs can vary depending on the specific model and task at hand. When it comes to how correlated the errors are, we observe high error correlation between LLMs and few-shot classifiers on known classes. From a bias–variance perspective (Valentini and Dietterich, 2002), this increased overlap on known classes may reflect reduced variance and shared biases across models, resulting in overlapping predictions and, by extension, correlated errors. This finding is consistent with the broader understanding that the phrasing of prompts can steer LLMs toward producing certain words more frequently (P. Liu et al., 2023), reinforcing shared biases across models and thereby contributing to correlated errors. In contrast, for unknown classes, where the prediction overlap is weaker, models exhibit different generalisation strategies, resulting in lower error correlation. Unlike in the case of known class classification, LLMs cannot rely solely on learned semantic patterns during training, since the unknown class is artificially introduced in our simulation study, and the concept of ”unknown” requires a more nuanced understanding than pure correlation. On the other hand, the traditional ML-based few-shot classifiers used in this work rely on the representation transfer and the geometric metric learned during supervised training for unknown class detection. These distinct inference mechanisms may explain why the error correlation for unknown class classification is weaker than for known class classification. While our current work does not demonstrate improved performance through model combination, the partial correlation of errors offers a promising direction for future research into hybrid or ensemble approaches.

With these findings, we have multiple theoretical and practical contributions. First, by introducing and comparing two different prompting strategies, we derive actionable recommendations for practitioners on which strategy is better for their use case. For instance, in fraud detection, precision may be considered more important than recall, as too many false positives may overwhelm investigators and reduce trust in the system (Wallny, 2022). On the other hand, if the costs of missing a fraudulent transaction are expensive, e.g. in high-value trading, it may be preferable to prioritise recall, even at the expense of generating more false positives. In this case, having a higher recall might be more beneficial. Moreover, we have also shown for implicit strategies that supplying examples may marginally increase classification performance, which additionally helps practitioners in designing their prompts. In terms of differing error patterns for unknown class detection, practitioners can exploit the differences to design more reliable systems. For instance, differing predictions could trigger human-in-the-loop review, or they could be combined into a confidence-weighted ensemble model for more accurate forecasting. On a more theoretical level, the error analysis and its comparison with existing few-shot classifiers provide a more nuanced understanding of the strengths and weaknesses of ICL.

Despite promising results, we acknowledge the limitations of our work that need further investigation and improvement. For instance, it is unclear whether the datasets used in this work have been used to train the LLMs. This data leakage is known to cause ”common token bias” (Zhao et al., 2021), which questions the generalisability of our findings. Additionally, the simplicity of the datasets used in this work may not reflect the real world in the most accurate manner (Arora et al., 2024). The relative brevity of the texts also raises questions about whether the results can be generalised to longer texts. Furthermore, given the dynamic nature of language models, our results represent a snapshot in time and will evolve, especially when newer architectures, such as text-oriented diffusion models (Lin et al., 2023), emerge.

Future work should investigate how to leverage the uncorrelated errors between traditional classifiers and large language models. For instance, more probabilistic approaches of quantifying the uncertainty of LLMs, such as semantic entropy (Farquhar et al., 2024) or SelfCheckGPT (Manakul et al., 2023), can be used further to investigate the differences between traditional classifiers and LLMs. Additionally, instead of randomly sampling OOD examples, other sampling strategies, such as kNN-based sampling, may also warrant additional insights into the decision-making process of ICL regarding unknown and known classes. Moreover, applying other prompting paradigms, such as ReAct (Yao et al., 2023) or reasoning-based approaches (Huang and Chang, 2023), could further validate our findings and shed light on their generalizability across prompting strategies.

6. Conclusion

Intent recognition, especially in the age of LLMs, is a crucial area impacting many business applications. Through empirical evaluation involving varying proportions of OOD examples across three datasets, we compare three existing few-shot classifiers with the ICL capabilities of four LLMs. In addition to providing recommendations for prompting strategies, our results support the general reliability of LLMs, but also reveal that for a significant number of cases, traditional few-shot classifiers and LLMs disagree on whether an example should be classified as unknown. These findings contribute to the broader discourse on the limitations and biases of LLMs and offer new insights for developing more robust, generalizable intent recognition systems. Given these results, future work should explore probabilistic and ensemble-based approaches that combine the complementary strengths of traditional classifiers and LLMs to improve performance on OOD examples in real-world applications.

References

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., Lee, J. R., Lee, Y. T., Li, Y., Liu, W., Mendes, C. C. T., Nguyen, A., Price, E., Rosa, G. d., Saarikivi, O., … Zhang, Y. (2024, December). Phi-4 Technical Report [arXiv:2412.08905 [cs]]. Retrieved June 2, 2025, from http://arxiv.org/abs/2412.08905

Anthropic. (2023). Welcome to claude. https://docs.anthropic.com/claude/docs/intro-to-claude

Arora, G., Jain, S., & Merugu, S. (2024, November). Intent detection in the age of LLMs. In F. Dernoncourt, D. Preoţiuc-Pietro, & A. Shimorina (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing: Industry track* (pp. 1559–1570). Association for Computational Linguistics.

Baier, L., Jöhren, F., & Seebacher, S. (2019). Challenges in the deployment and operation of machine learning in practice. *Proceedings of the 27th european conference on information systems (ECIS)*.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, *24*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Casanueva, I., Temčinas, T., Gerz, D., Henderson, M., & Vulić, I. (2020). Efficient Intent Detection with Dual Sentence Encoders. *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, 38–45.

Chen, G., Xu, Q., Zhan, C., Wang, F. L., Liu, K., Liu, H., & Hao, T. (2024). Improving open intent detection via triplet-contrastive learning and adaptive boundary. *IEEE Transactions on Consumer Electronics*, *70*(1), 2806–2816.

Chen, J., Zhang, R., Jiang, X., & Hu, C. (2024). SPContrastNet: A Self-Paced Contrastive Learning Model for Few-Shot Text Classification. *ACM Transactions on Information Systems*, *42*(5), 1–25.

Chen, J., Zhang, R., Mao, Y., & Xu, J. (2022). Contrastnet: A contrastive learning framework for few-shot text classification [Number: 10]. *Proceedings of the AAAI conference on artificial intelligence*, *36*, 10492–10500.

Chen, M., Du, J., Pasunuru, R., Mihaylov, T., Iyer, S., Stoyanov, V., & Kozareva, Z. (2022). Improving In-Context Few-Shot Learning via Self-Supervised Training. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3558–3573.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (0th ed.). Routledge. https://doi.org/10.4324/9780203771587

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., & Sui, Z. (2024). A Survey on In-context Learning. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1107–1128.

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, *630*(8017), 625–630.

Gemini Team. (2023). Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Geng, C., Huang, S.-J., & Chen, S. (2021). Recent Advances in Open Set Recognition: A Survey [Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(10), 3614–3631.

Gu, Y., Dong, L., Wei, F., & Huang, M. (2023). Pre-Training to Learn in Context. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4849–4870.

He, M., & Garner, P. N. (2023). Can ChatGPT Detect Intent? Evaluating Large Language Models for Spoken Language Understanding. *INTERSPEECH 2023*, 1109–1113.

Honovich, O., Shaham, U., Bowman, S. R., & Levy, O. (2023). Instruction Induction: From Few Examples to Natural Language Task Descriptions. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1935–1952.

Huang, J., & Chang, K. C.-C. (2023). Towards Reasoning in Large Language Models: A Survey. *Findings of the Association for Computational Linguistics: ACL 2023*, 1049–1065.

Jin, Q., Wan, N., Leaman, R., Tian, S., Wang, Z., Yang, Y., Wang, Z., Xiong, G., Lai, P.-T., Zhu, Q., et al. (2024). Demystifying Large Language Models for Medicine: A Primer. *ArXiv*, arXiv–2410.

Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al. (2025). Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Kerby, D. S. (2014). The Simple Difference Formula: An Approach to Teaching Nonparametric Correlation. *Comprehensive Psychology*, *3*, 11.IT.3.1. https://doi.org/10.2466/11.IT.3.1

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners [event-place: New Orleans, LA, USA]. *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

Kossen, J., Gal, Y., & Rainforth, T. (2023). In-context learning learns label relationships but is not conventional learning. *International conference on learning representations*.

Kuligowska, K., & Kowalczuk, B. (2024). Dataset Expansion with Pseudo-Labeling: Case Study for Optimizing Chatbot Intent Recognition. *Humanities & Social Sciences Reviews*, *12*(2), 104–109.

Kumar, D., AbuHashem, Y. A., & Durumeric, Z. (2024). Watch your language: Investigating content moderation with large language models. *Proceedings of the international AAAI conference on web and social media*, *18*, 865–878.

Lang, H., Zheng, Y., Hui, B., Huang, F., & Li, Y. (2024, May). Out-of-domain intent detection considering multi-turn dialogue contexts. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)* (pp. 12539–12552). ELRA; ICCL.

Larson, S., Mahendran, A., Peper, J. J., Clarke, C., Lee, A., Hill, P., Kummerfeld, J. K., Leach, K., Laurenzano, M. A., Tang, L., & Mars, J. (2019). An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1311–1316.

Li, S., Song, Z., Xia, Y., Yu, T., & Zhou, T. (2024). The closeness of in-context learning and weight shifting for softmax regression. *Advances in Neural Information Processing Systems*, *37*, 62584–62616.

Li, X., & Li, J. (2024, August). AoE: Angle-optimized embeddings for semantic textual similarity. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1825–1839). Association for Computational Linguistics.

Lin, Z., Gong, Y., Shen, Y., Wu, T., Fan, Z., Lin, C., Duan, N., & Chen, W. (2023). Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise. *International conference on machine learning*, 21051–21064.

Liu, B., Zhan, L.-M., Lu, Z., Feng, Y., Xue, L., & Wu, X.-M. (2024, May). How good are LLMs at out-of-distribution detection? In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)* (pp. 8211–8222). ELRA; ICCL.

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., & Chen, W. (2022). What Makes Good In-Context Examples for GPT-3? *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, 100–114.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, *55*(9), 1–35. https://doi.org/10.1145/3560815

Liu, X., Eshghi, A., Swietojanski, P., & Rieser, V. (2021). Benchmarking natural language understanding services for building conversational agents. *Increasing naturalness and flexibility in spoken dialogue interaction: 10th international workshop on spoken dialogue systems*, 165–183.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022, May). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 8086–8098). Association for Computational Linguistics.

Manakul, P., Liusie, A., & Gales, M. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *Proceedings of the 2023 conference on empirical methods in natural language processing*, 9004–9017.

Meta AI. (2024a). LLaMA 3.1 Model Card. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1/

Meta AI. (2024b). Llama 3.3 Model Card. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/

Min, S., Lewis, M., Zettlemoyer, L., & Hajishirzi, H. (2022). MetaICL: Learning to Learn In Context. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2791–2809.

Mina, M., Ruíz-Fernández, V., Falcão, J., Vasquez-Reina, L., & González-Agirre, A. (2025). Cognitive biases, task complexity, and result intepretability in large language models. *Proceedings of the 31st international conference on computational linguistics*, 1767–1784.

Oliver, M., & Wang, G. (2024). Crafting Efficient Fine-Tuning Strategies for Large Language Models [Version Number: 1]. Retrieved January 16, 2025, from https://arxiv.org/abs/2407.13906

OpenAI. (2022). Introducing ChatGPT. Retrieved May 5, 2025, from https://openai.com/index/chatgpt/

Parikh, S., Tiwari, M., Tumbade, P., & Vohra, Q. (2023). Exploring Zero and Few-shot Techniques for Intent Classification. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, 744–751.

Ran, Z., Chen, Y., Jiang, Q., & E, K. (2024). Intent Recognition in Dialogue Systems. *Proceedings of the 4th Asia-Pacific Artificial Intelligence and Big Data Forum*, 165–173.

Shi, W., Min, S., Lomeli, M., Zhou, C., Li, M., Szilvasy, G., James, R., Lin, X. V., Smith, N. A., Zettlemoyer, L., et al. (2023). In-context Pretraining: Language Modeling Beyond Document Boundaries. *arXiv e-prints*, arXiv–2310.

Sidlauskiene, J., Joye, Y., & Auruskeviciene, V. (2023). AI-based chatbots in conversational commerce and their effects on product and price perceptions. *Electronic Markets*, *33*(1), 24.

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Sun, J., Mei, C., Wei, L., Zheng, K., Liu, N., Cui, M., & Li, T. (2024). Dial-insight: Fine-tuning Large Language Models with High-Quality Domain-Specific Data

Preventing Capability Collapse. *arXiv preprint arXiv:2403.0 9167*.

Sun, L.-Y., & Chu, W.-T. (2024). Overall positive prototype for few-shot open-set recognition. *Pattern Recognition*, *151*, 110400.

Suryanto, T. L. M., Wibawa, A. P., Hariyono, H., & Nafalski, A. (2023). Evolving Conversations: A Review of Chatbots and Implications in Natural Language Processing for Cultural Heritage Ecosystems. *International Journal of Robotics and Control Systems*, *3*(4), 955–1006.

Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.

Valentini, G., & Dietterich, T. G. (2002). Bias—Variance Analysis and Ensembles of SVM [Series Title: Lecture Notes in Computer Science]. In G. Goos, J. Hartmanis, J. Van Leeuwen, F. Roli, & J. Kittler (Eds.), *Multiple Classifier Systems* (pp. 222–231, Vol. 2364). Springer Berlin Heidelberg.

Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu koray, k., & Wierstra, D. (2016). Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29). Curran Associates, Inc.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., & Vladymyrov, M. (2023, July). Transformers learn in-context by gradient descent. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th international conference on machine learning* (pp. 35151–35174, Vol. 202). PMLR.

Wallny, F. (2022). False Positives in Credit Card Fraud Detection: Measurement and Mitigation. https://doi.org/10.24251/HICSS.2022.195

Wang, P., He, K., Wang, Y., Song, X., Mou, Y., Wang, J., Xian, Y., Cai, X., & Xu, W. (2024, May). Beyond the known: Investigating LLMs performance on out-of-domain intent detection. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)* (pp. 2354–2364). ELRA; ICCL.

Wang, X., Wang, Y., Xu, C., Geng, X., Zhang, B., Tao, C., Rudzicz, F., Mercer, R. E., & Jiang, D. (2023, August). Investigating the Learning Behaviour of In-context Learning: A Comparison with Supervised Learning [arXiv:2307.15411 [cs]]. Retrieved November 4, 2024, from http://arxiv.org/abs/2307.15411

Wang, Y., Chao, W.-L., Weinberger, K. Q., & Maaten, L. v. d. (2019, November). SimpleShot: Revisiting Nearest-Neighbor Classification for Few-Shot Learning [arXiv:1911.04623 [cs]]. Retrieved April 30, 2025, from http://arxiv.org/abs/1911.04623

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models [event-place: New Orleans, LA, USA]. *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

Wei, J., Hou, L., Lampinen, A., Chen, X., Huang, D., Tay, Y., Chen, X., Lu, Y., Zhou, D., Ma, T., et al. (2023). Symbol tuning improves in-context learning in language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 968–979.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). React: Synergizing reasoning and acting in language models. *International Conference on Learning Representations (ICLR)*.

Yehudai, A., & Bandel, E. (2024). FastFit: Fast and Effective Few-Shot Text Classification with a Multitude of Classes. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, 174–184.

Zhang, H., Zhang, Y., Zhan, L.-M., Chen, J., Shi, G., Wu, X.-M., & Lam, A. Y. (2021). Effectiveness of Pre-training for Few-shot Intent Classification. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1114–1120.

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. *International conference on machine learning*, 12697–12706.

Zhu, Z., Cheng, X., An, H., Wang, Z., Chen, D., & Huang, Z. (2024, May). Zero-shot spoken language understanding via large language models: A preliminary study. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)* (pp. 17877–17883). ELRA; ICCL.